Data Analysis for HPC Resilience: A Perspective from Statistics

George Ostrouchov

Statistics and Data Sciences Group
Computer Science and Mathematics Division
Oak Ridge National Laboratory

2009 National HPC Workshop on Resilience
August 13-14, 2009
Washington, DC

This work was partially funded by the Department of Defense and the Extreme Scale Systems Center at ORNL

Collaborators

Stephen L. Scott

Christian Engelmann

Geoffroy Vallee

Thomas Naughton

System Research Team
Computer Science Research Group
Computer Science and Mathematics Division
Oak Ridge National Laboratory

Chokchai (Box) Leangsuksun

Associate Professor of Computer Science
The SWEPCO Endowed Professor
Center for Entrepreneurship and Information Technology
Louisiana Tech University

Stephen W. Poole

Chief Scientist, Computer Science and Mathematics Division
LCF System Architect, National Center for Computational Sciences
Director, Extreme Scale Systems Center
Oak Ridge National Laboratory

JSM 2009 Washington DC

Statistics: from Evidence to Policy

JSM 2009 ONLINE PRELIMINARY PROGRAM

Add To My Program

245 **Tue, 8/4/09, 8:30 AM - 10:20 AM** **CC-153**

Advanced Reliability Methods with Applications - Invited - Papers

Section on Physical and Engineering Sciences, Section on Quality and Productivity

Organizer(s): I-Li Lu, The Boeing Company

Chair(s): Ranjan Paul, The Boeing Company

8:35 AM	**Using Accelerated Life Tests Results to Predict Product Field Reliability** — William Q. Meeker, Iowa State University; Luis A. Escobar, Louisiana State University; Yili Hong, Iowa State University
9:00 AM	**The Development of Advanced Reliability Methods for Aircraft Maintenance Optimization Process** — I-Li Lu, The Boeing Company; Anbessie A. Yitbarek, The Boeing Company; Ranjan Paul, The Boeing Company; Elizabeth A. Whalen, The Boeing Company; Shuying Zhu, The Boeing Company
9:25 AM	**Reliability in Supercomputing: A Million Processors Cooperating to Solve One Problem** — George Ostrouchov, Oak Ridge National Laboratory; Thomas J. Naughton, III, Oak Ridge National Laboratory; Stephen L. Scott, Oak Ridge National Laboratory
9:50 AM	**Detection of Nuclear Material Entering Ports: An Analytic Framework for Data and Policy Analysis** — Siddhartha Dalal, RAND Corporation
10:15 AM	Floor Discussion

JSM2009 Washington DC

Statistics: from Evidence to Policy

JSM 2009 ONLINE PRELIMINARY PROGRAM

Add To My Program

245 **Tue, 8/4/09, 8:30 AM - 10:20 AM** **CC-153**

Advanced Reliability Methods with Applications - Invited - Papers

Section on Physical and Engineering Sciences, Section on Quality and Productivity

Organizer(s): I-Li Lu, The Boeing Company

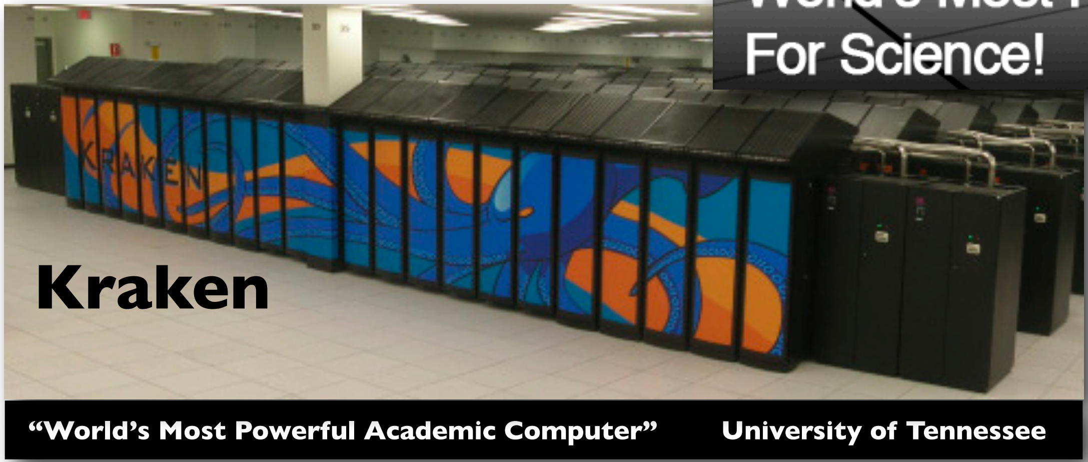
Chair(s): Ranjan Paul, The Boeing Company

8:35 AM	**Using Accelerated Life Tests Results to Predict Product Field Reliability** — William Q. Meeker, Iowa State University; Luis A. Escobar, Louisiana State University; Yili Hong, Iowa State University
9:00 AM	**The Development of Advanced Reliability Methods for Aircraft Maintenance Optimization Process** — I-Li Lu, The Boeing Company; Anbessie A. Yitbarek, The Boeing Company; Ranjan Paul, The Boeing Company; Elizabeth A. Whalen, The Boeing Company; Shuying Zhu, The Boeing Company
9:25 AM	**Reliability in Supercomputing: A Million Processors Cooperating to Solve One Problem** — George Ostrouchov, Oak Ridge National Laboratory; Thomas J. Naughton, III, Oak Ridge National Laboratory; Stephen L. Scott, Oak Ridge National Laboratory
9:50 AM	**Detection of Nuclear Material Entering Ports: An Analytic Framework for Data and Policy Analysis** — Siddhartha Dalal, RAND Corporation
10:15 AM	Floor Discussion

Equipment for Computational Experiments: Component Count is Increasing

top500.org processor count:

- Less than three years ago the entire 500 list broke the million processor mark
- Now the top 6 add up to over a million



Help! - There is Only One Statistician in the Machine Room!

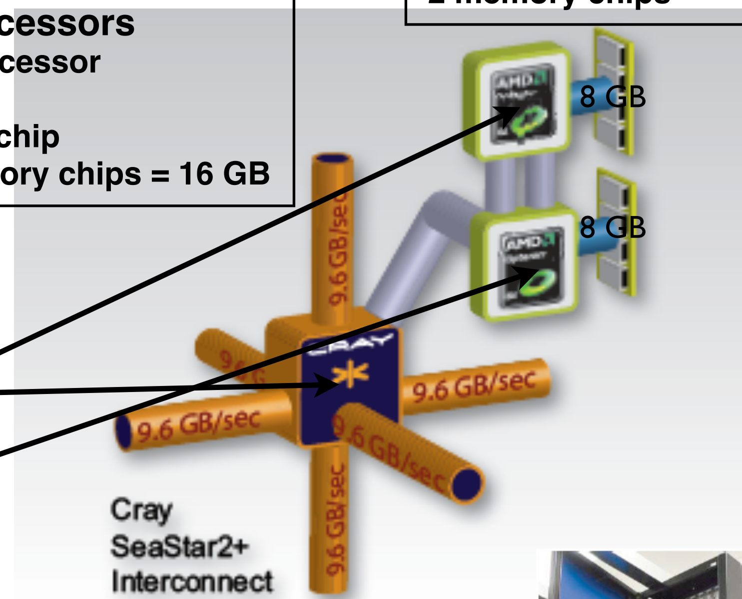
Cray XT5: Over 1,400 Components Packed Into Each Cabinet

Processor = 4 Cores
2 memory chips

Node = 2 Processors
4 cores per processor

1 interconnect chip
4 x (4 GB) memory chips = 16 GB

Blade = 4 Nodes
8 processors
32 cores
4 interconnect chips
16 (4 GB) memory chips = 64 GB
6 DC voltage converters



Cabinet = 24 Blades
768 cores
96 interconnect chips
384 memory chips (1.5 TB)
144 voltage converters
+ power supply, liquid cooling, etc.
Power 480V, ~40,000 Watt per cabinet

Jaguar = 284 cabinets (XT5 and XT4), ~ 6.5 Megawatts

Reactive and Proactive Fault Tolerance

- Reactive: Keeps applications alive through recovery from experienced failures
  - Checkpoint/restart
  - Message logging/replay
  - Effective until failures get too frequent
  - Timely failure reporting for restart
  - Root cause analysis for repair
- Proactive: Keeps applications alive by avoiding failures
  - Preemptive migration
  - Effectively decreases failure rate
  - Needs failure prediction

Reactive and Proactive Fault Tolerance

- **Reactive: Keeps applications alive through recovery from experienced failures**
  - Checkpoint/restart
  - Message logging/replay
  - Effective until failures get too frequent
  - Timely failure reporting for restart
  - Root cause analysis for repair
- **Proactive: Keeps applications alive by avoiding failures**
  - Preemptive migration
  - Effectively decreases failure rate
  - Needs failure prediction

Data Analysis

George Ostrouchov - ostrouchovg@ornl.gov

HPC Data: Solutions Still Evolving

- **Log file data**
  - Automated record of system events
  - Important events are buried in unimportant events and in countless duplicates triggered in other system components.
  - Time-dependent component failure data (record of component replacements)
- **System state data**
  - Measurements collected actively or passively, at regular intervals, about hardware states and software states.
- **System structure data**
  - Hardware dependencies, system software dependencies, application software dependencies
  - Initial component composition
  - "Catalyst" for building predictive models.

Analysis is First Focused on Anomalies

- **The few that are different from most**
  - Different from neighbors
  - Smallest clusters
  - Identify unusual by estimating usual
  - Quality Control Methods
  - Fraud detection ideas (Credit card and telephone industries)
  - Action needs further decision on anomaly type
- **Identify specific anomalies**
  - Pareto diagram (most frequent first)
  - Action may be unique
- **Automated selection of features that are relevant**
  - for what?
  - Variable selection

Attributes:
Sampled value
Average value
Variance
Histogram
Density estimate
Posterior density
Regression parameter
Term frequency count

	A1	A2	B1	B2	B3	...
n1h1						
n1h2						
. . .						
n2h1						
n2h2						
. . .						
. . .						

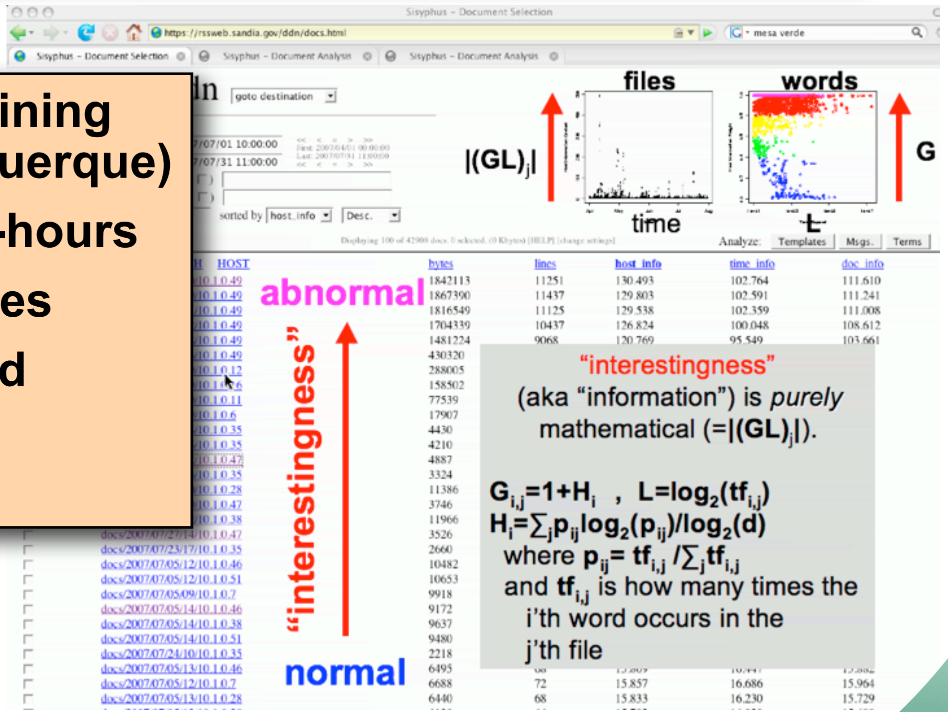
Items:
Node hours
Node minutes
Processor seconds
voltage converter minutes

Log File Data

- Enormous files with many events from all nodes
  - A single event may trigger events on hundreds of other nodes
  - Usually clocks not synchronized
  - Determining root cause very difficult
  - Used for system MTTF estimation after extensive filtering

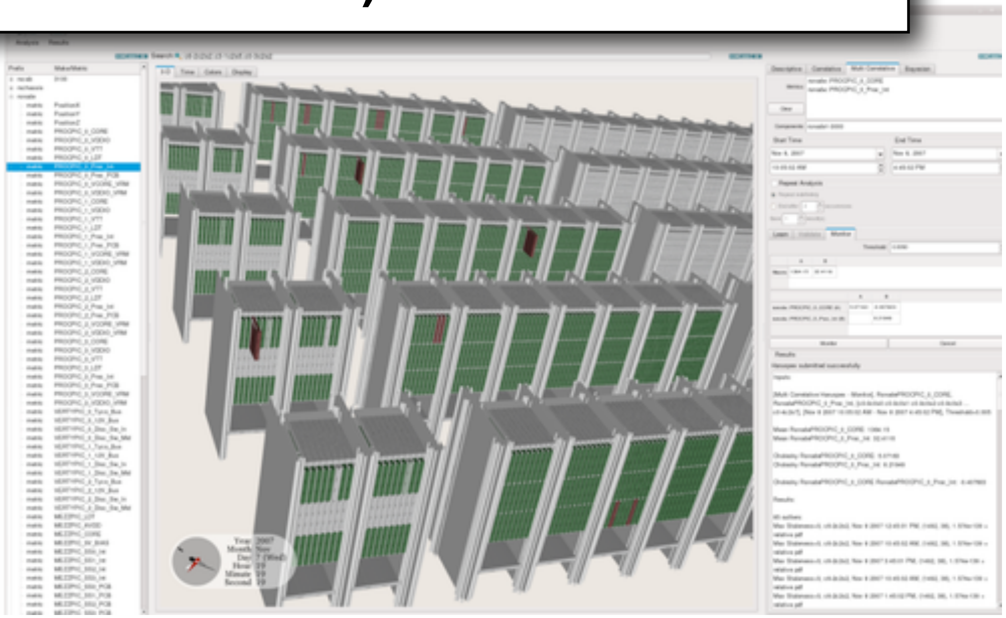
Sysiphus: an event-log data mining toolkit (Stearley, Sandia-Albuquerque)

- Divide messages into node-hours
- Term frequencies are features
- Use text-mining ideas to find anomalous node-hours



System State Data

- **Ganglia: a scalable distributed monitoring system**
  - **RRDtool (round robin data)**
- **OVIS: A Tool for Intelligent, Scalable, Real-Time Monitoring of Large Computational Clusters (Brandt et al., Sandia-Livermore)**



Active Data Collection: Measuring Operating System Noise

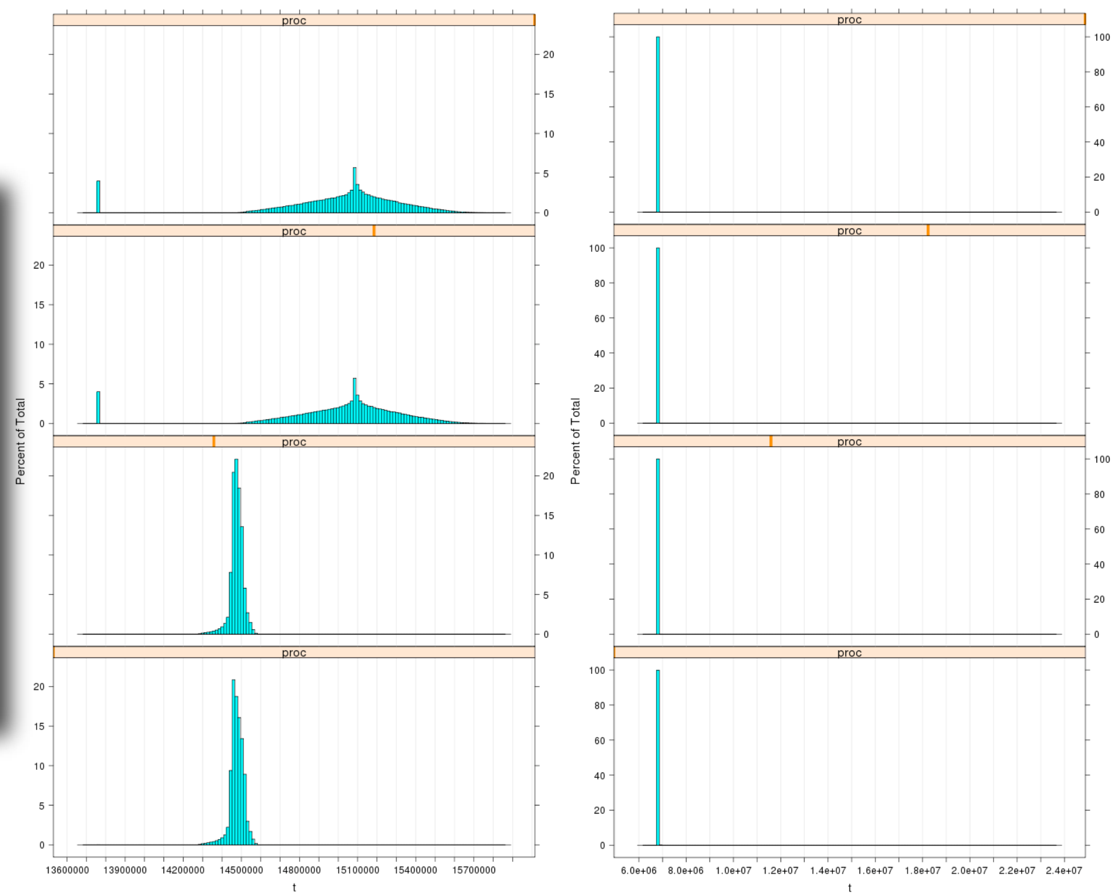
FWQ (fixed work quantity): Amount of time to run a fixed amount of computation

FTQ (fixed time quantity): Amount of work completed in fixed amount of time

Sensitive to operating system interrupts

May be another system state measured attribute

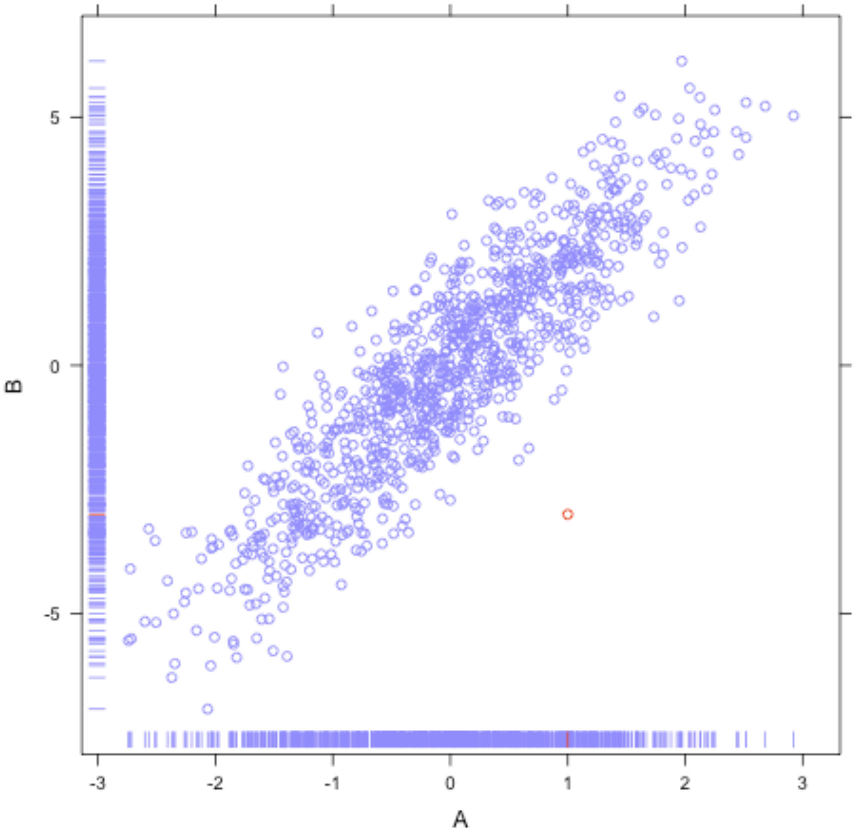
OS: built from ground up | **OS: Linux pared down**



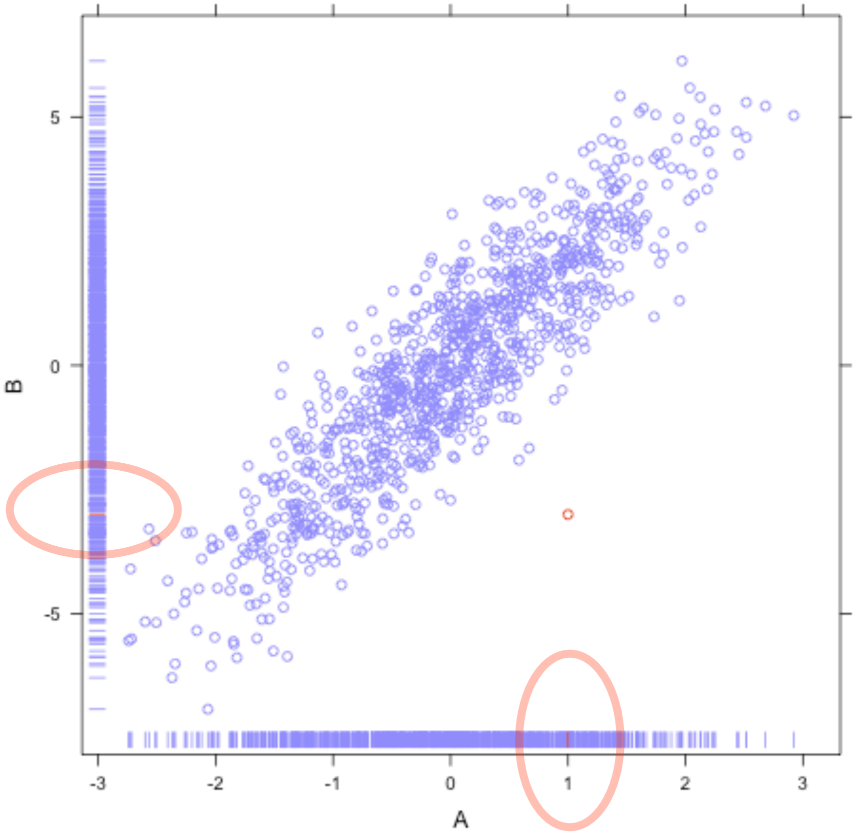
13.5 | 15.7 | 6.0 | 24.0

Coefficient of variation ~0.03 | **Coefficient of variation ~1.2**

Why do Multivariate Methods Matter?

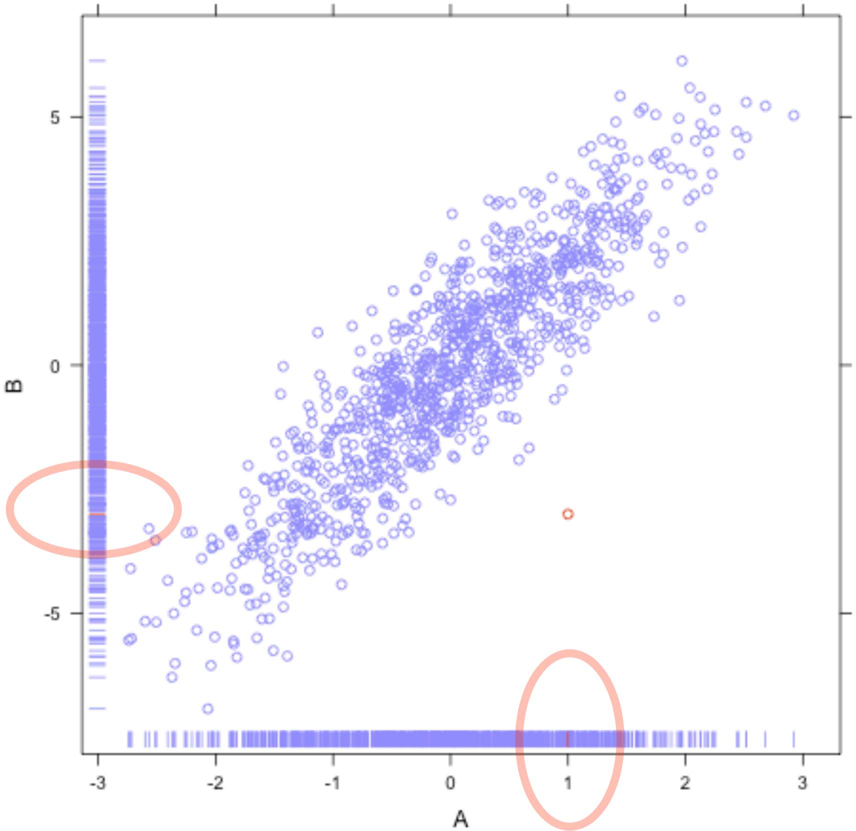


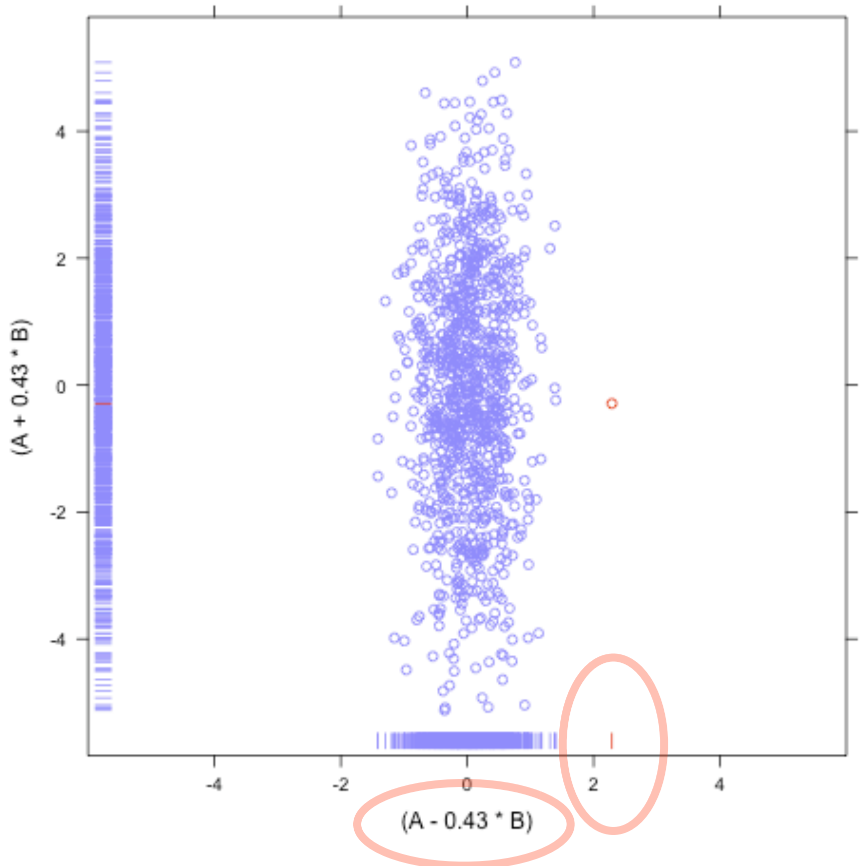
Why do Multivariate Methods Matter?



- **Outlier not visible by univariate methods**

Why do Multivariate Methods Matter?

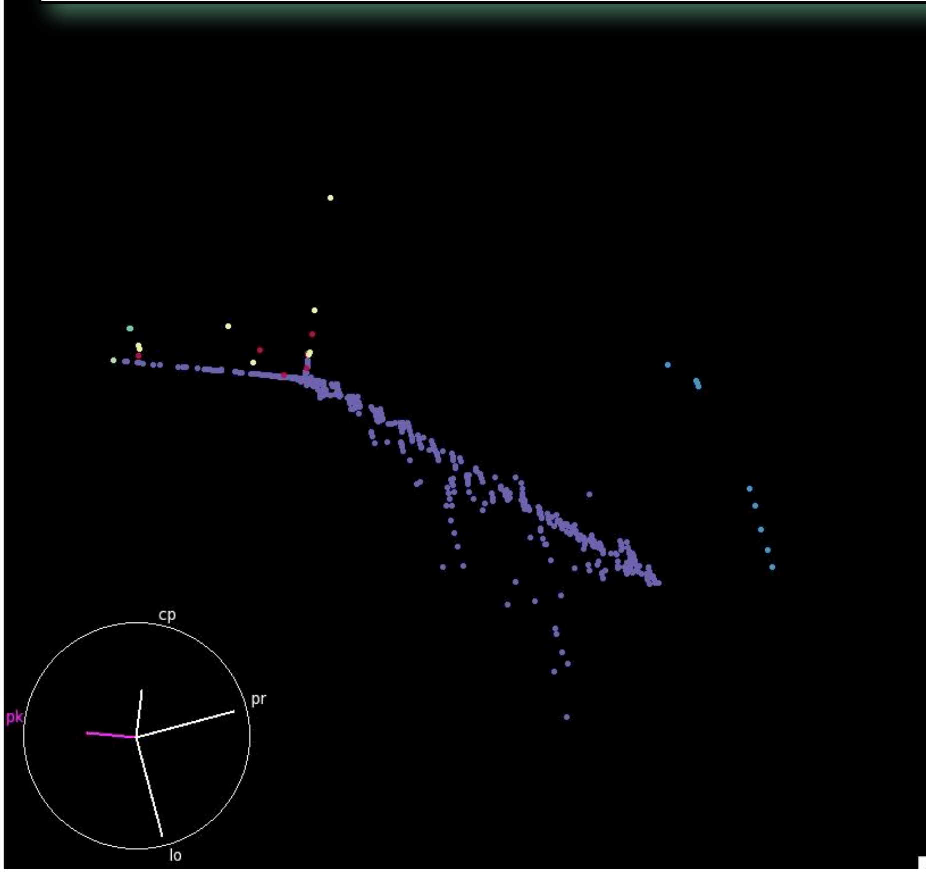


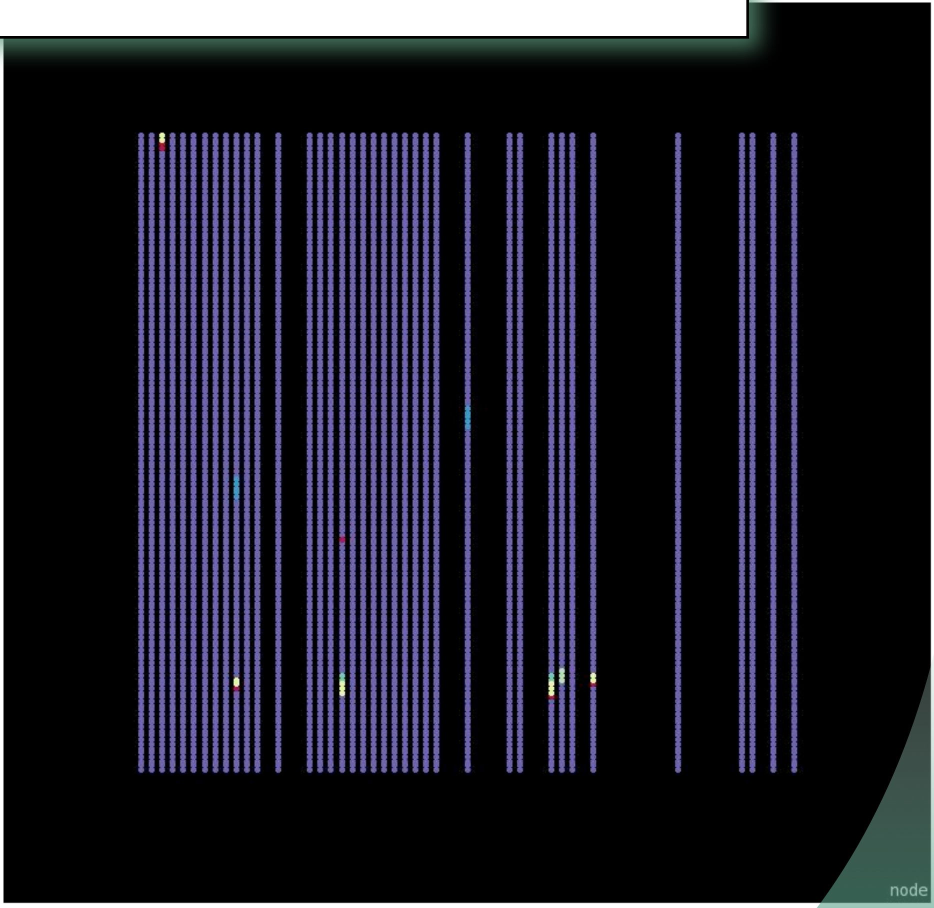


- **Outlier not visible by univariate methods**
- **Know the right feature to compute**

Injecting Faults on an Old Cluster

- **Clustering node-minute ganglia data (R/GGobi)**
- **Anomaly measure (purple-blue-yellow-red) proportional to cluster size**
- **Found all injected faults**





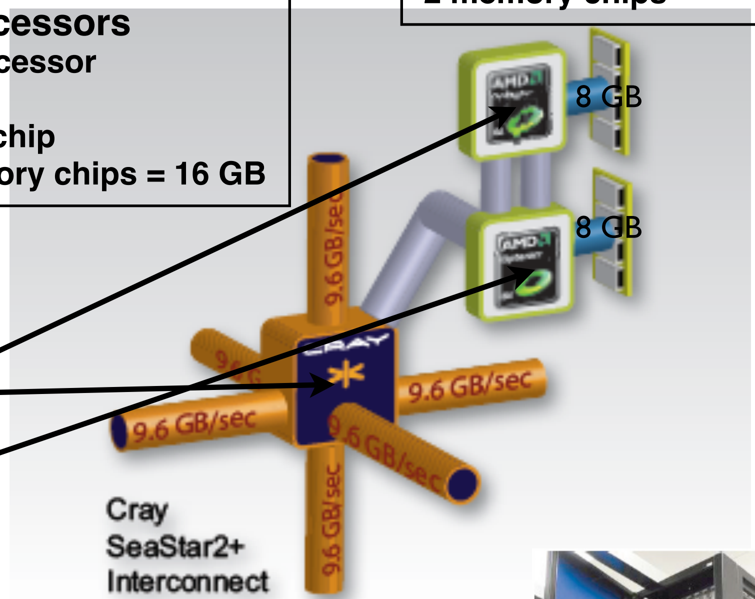
Identifying anomalies: first step to identifying failures and building a failure prediction capability

Cray XT5: Over 1,400 Components Packed Into Each Cabinet

Processor = 4 Cores
2 memory chips

Node = 2 Processors
4 cores per processor

1 interconnect chip
4 x (4 GB) memory chips = 16 GB



Blade = 4 Nodes
8 processors
32 cores
4 interconnect chips
16 (4 GB) memory chips = 64 GB
6 DC voltage converters

Cabinet = 24 Blades
768 cores
96 interconnect chips
384 memory chips (1.5 TB)
144 voltage converters
+ power supply, liquid cooling, etc.
Power 480V, ~40,000 Watt per cabinet

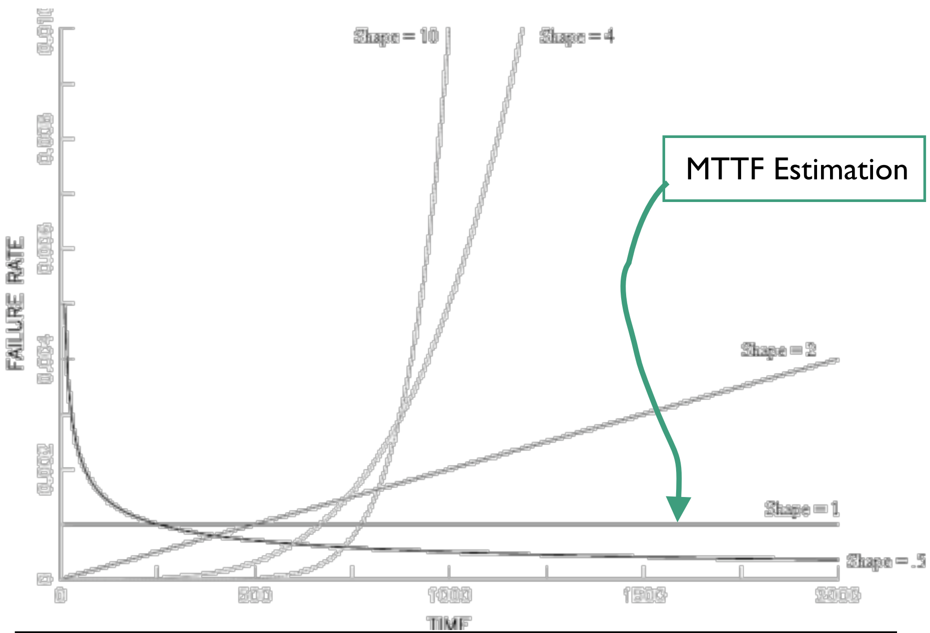
Jaguar = 284 cabinets (XT5 and XT4), ~ 6.5 Megawatts

HPC: Amazing Laboratory for Testing Hardware and Software Components

- Many replicates of each hardware type on test
- Many test situations for each piece of system software
- Need time-dependent component reliability data to move beyond MTTF estimation
  - System component inventory
  - Component replacement event record including cause
  - Helps root cause analysis
- Same failure rate after replacement indicates external cause
  - The power of replicates
- Software reliability
- Accelerated testing (temperature? overclocking??)

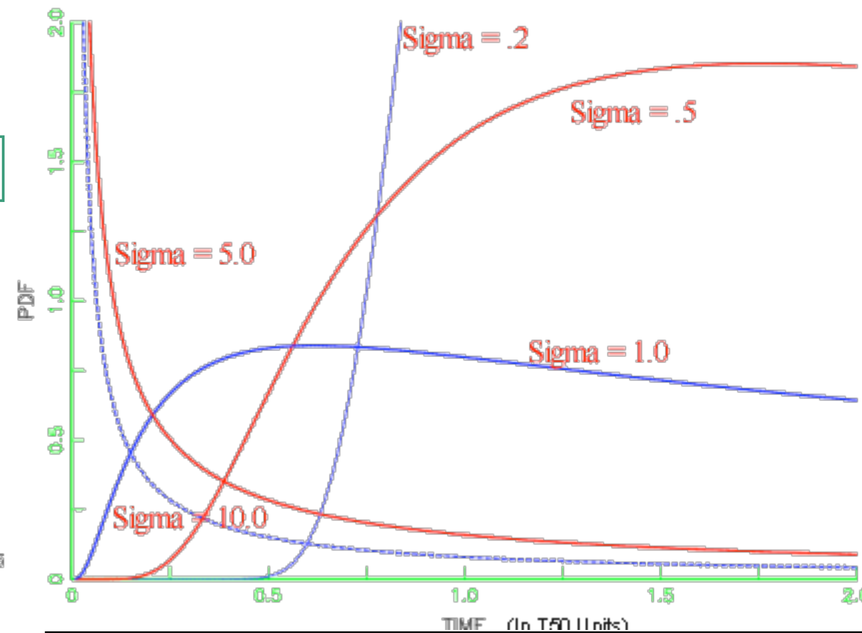
Statistical Reliability Models

- T: time of failure
- Lifetime distribution function: P(T ≤ t)
- Lifetime density function: d/dt P(T ≤ t)
- Survival function: P(T > t)
- Hazard function: failure rate at time t "bathtub curve"
- Many lifetime distributions
- Mixtures often used for multiple failure modes



Weibull Hazard Shapes

Failure of the "weakest link" of many competing failure processes



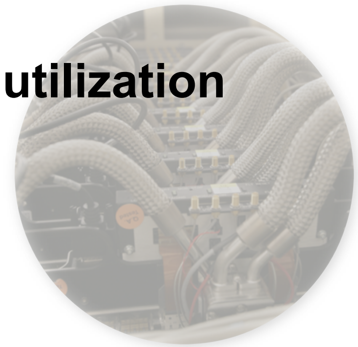
Lognormal Hazard Shapes

Failures resulting from incremental degradation processes

Source: NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/, 2009

Are HPC Reliability Issues Different from Other Reliability Applications?

- **Characteristics**
  - Lots of replication to find outliers
  - Anything (almost) can be sampled or computed
  - Anything must have low (or background) resource utilization
  - Centralized solution does not scale
  - Fault tolerant estimation algorithms
  - Short life cycle of HPC systems



- **What can statistics bring to HPC Reliability?**
  - Survival Analysis and Reliability Analysis methods
  - Quality Control Methods
  - Multivariate data analysis
  - Probability based models (Likelihood, Bayesian, etc.)
  - Design of experiments